Current Biology

A single community dominates structure and function of a mixture of multiple methanogenic communities

--Manuscript Draft--

Manuscript Number:	CURRENT-BIOLOGY-D-17-00432R4
Full Title:	A single community dominates structure and function of a mixture of multiple methanogenic communities
Article Type:	Report
Corresponding Author:	Pawel Sierocinski University of Exeter Penryn, Cornwall UNITED KINGDOM
First Author:	Pawel Sierocinski
Order of Authors:	Pawel Sierocinski
	Kim Milferstedt
	Florian Bayer
	Tobias Großkopf
	Mark Alston
	Sarah Bastkowski
	David Swarbreck
	Phil J Hobbs
	Orkun S Soyer
	Jérôme Hamelin
	Angus Buckling
Abstract:	The ecology of microbes frequently involves the mixing of entire communities (community coalescence), for example flooding events, host excretion and soil tillage [1,2], yet the consequences of this process for community structure and function are poorly understood [3-7]. Recent theory suggests that a community, due to coevolution between constituent species, may act as a partially cohesive unit [8-11], resulting in one community dominating following community coalescence. This dominant community is predicted to be the one that uses resources most efficiently when grown in isolation [11]. We experimentally tested these predictions using methanogenic communities, for which efficient resource use, quantified by methane production, requires coevolved cross-feeding interactions between species [12]. Following propagation in laboratory-scale anaerobic digesters, community composition (determined from 16S rRNA sequencing) and methane production of mixtures of communities closely resembled that of the single most productive community grown in isolation. Analysis of each community's contribution towards the final mixture suggests that certain combinations of taxa within a community might be co-selected as a result of coevolved interactions. As a corollary of these findings, we also show that methane production increased with the number of inoculated communities. These findings are relevant to the understanding of the ecological dynamics of natural microbial communities, as well as demonstrating a simple method of predictably enhancing microbial community function in biotechnology, health and agriculture [13].

Dear Christine,

Many thanks for your email and the suggested corrections. They were very clear so hopefully now everything is up to standard. Please, could you also thank the reviewers for their contributions? We honestly appreciate their comments as they have led to a greatly improved manuscript compared with the initial version. And most importantly it should be a much clearer read in its current form.

Best wishes,

Pawel Sierocinski & Angus Buckling

"I think the authors need to clarify what is a community-level process versus a process driven by selection on individual species. Let's suppose that some syntrophic metabolic feature within a methane-producing community is driven to increase in frequency (whether by community or individual selection), such that all species contributing to the particular metabolic feature increase in frequency together. For example, we might imagine that acetogenic members and the species most closely interacting with them increase in frequency as a consequence of their syntrophies. If this is the case, two possibilities exist. First, the various species may increase together because of many complex interacting features such that no other species would succeed together. I would call this a community-level process. Alternatively, the acetogenic members may increase just because they produce acetate and that any acetogenic species would equally succeed. It seems to me that if there is interchangeability of species that would increase together, then this reflects species-level selection and not community-level selection. It would generally be hard to rule out species-level selection"

1. When discussing the role of "community level properties" in explaining a community's contribution to the mixture, we mean that specific coevolved interactions are important for determining an organism's fitness, such that the presence of one taxon enhances the fitness of another taxon. This probably has nothing to do with higher order or "community level selection", which theoretically has been shown to operate under very limited conditions, but instead is more parsimoniously explained by individual and co-selection: individual selection for a particularly successful genotype will in turn co-select for the organisms it interacts with. This then results in interacting taxa increasing (or decreasing) in frequency as modules. We therefore use the term co-selection instead of community level properties to avoid this confusion
2. We entirely agree with the reviewer's point of inter-changeability of species between communities: this is central to where or not co-selection has a role within communities. We make this more explicit
3. We further spell out the simple predictions regarding the relationship between productivity and community contribution, by giving an explicit illustration. When co-selection occurs for all taxa within a community and therefore a single community entirely dominates the mixture, no other community will contribute regardless of productivity.
4. We include both methane and cell density as measures of productivity, in response to the reviewer's previous concern that methane production alone may not be informative enough.

Your MS Word document "conflict-of-interest.doc" cannot be opened and processed. Please see the common list of problems, and suggested resolutions below.

Common Problems When Creating a PDF from Microsoft Word Documents

When you open your document in MS Word, an alert box may appear with a message. This message may relate to margins or document size. You will need to find the piece of your Word document that is causing the problem. Selectively remove various pieces of the file, saving the modified file with a temporary file name. Then try to open modified file. Repeat this process until the alert box no longer appears when you open the document.

Embedded Macros

Your submission should not contain macros. If they do, an alert box may appear when you open your document (this alert box prevents EM from automatically converting your Word document into the PDF that Editors and Reviewers will use). You must adjust your Word document to remove these macros.

Corrupted Tables

Your document may contain a table that cannot be rendered correctly. This will be indicated by a warning alert box. Correct the content of the table that causes the problem, so that the alert box no longer appears.

Word 2002/Word XP files

At the present time, EM supports Word files in Word 2000 and earlier formats. If you are using a more recent version of MS Word, try saving your Word document in a format compatible with Word 2000, and resubmit to EM.

Other Problems

If you are able to get your Word document to open with no alert boxes appearing, and you have submitted it in Word 2000 (or earlier) format, and you still see an error indication in your PDF file (where your Word document should be appearing). please contact the journal via the 'Contact Us' button on the Navigation Bar.'

You will need to reformat your Word document, and then re-submit it.

A single community dominates structure and function of a mixture of multiple methanogenic communities

Pawel Sierocinski*[1,6], Kim Milferstedt[2], Florian Bayer[1], Tobias Großkopf[3], Mark Alston[4], Sarah Bastkowski[4], David Swarbreck[4], Phil J Hobbs[5], Orkun S Soyer[3], Jérôme Hamelin[2], Angus Buckling[1]

[1] *Biosciences, University of Exeter, Penryn, Cornwall, TR10 9FE, UK*

[2] *LBE, INRA, 11100, Narbonne, France*

[3] *School of Life Sciences, University of Warwick, Coventry,* CV4 7AL*, UK*

[4] *Earlham Institute, Norwich Research Park, Norwich, NR4 7UH, UK*

[5] *Anaerobic Analytics Ltd, Okehampton, EX20 1AS, UK*

[6] Lead contact

*Correspondence should be directed to p.sierocinski@exeter.ac.uk

SUMMARY

The ecology of microbes frequently involves the mixing of entire communities (community coalescence), for example flooding events, host excretion and soil tillage [1,2]**, yet the consequences of this process for community structure and function are poorly understood** [3–7]**. Recent theory suggests that a community, due to coevolution between constituent species, may act as a partially cohesive unit** [8–11]**, resulting in one community dominating following community coalescence. This dominant community is predicted to be the one that uses resources most efficiently when grown in isolation** [11]**. We experimentally tested these predictions using methanogenic communities, for which efficient resource use, quantified by methane production, requires coevolved cross-feeding interactions between species** [12]**. Following propagation in laboratory-scale anaerobic digesters, community composition (determined from 16S rRNA sequencing) and methane production of mixtures of communities closely resembled that of the single most productive community grown in isolation. Analysis of each community's contribution towards the final mixture suggests that certain combinations of taxa within a community might be co-selected as a result of coevolved interactions. As a corollary of these findings, we also show that methane production increased with the number of inoculated communities. These findings are relevant to the understanding of the ecological dynamics of natural**

microbial communities, as well as demonstrating a simple method of predictably enhancing microbial community function in biotechnology, health and agriculture [13]**.**

RESULTS AND DISCUSSION

We wanted to determine if coalesced methanogenic communities were dominated by the community that used resources most efficiently in isolation. We used methanogenic communities primarily because methane production is a useful proxy for the ability of an anaerobic community to fully exploit available resources: Methanogenesis results from the conversion of H_2, CO_2 and short chain fatty acids produced by hydrolysis and fermentation of more complex organic material, and is often the only thermodynamically feasible way of actively removing inhibitory end-metabolites [12]. Moreover, methanogenic communities are characterized by complex cross-feeding interactions [12, 14, 15]; hence, the importance of community cohesion in shaping community performance is likely to be particularly important [9]. To provide insight into the temporal dynamics of compositional and functional change following community mixing, we first measured the methane production and composition of two methanogenic communities derived from industrial Anaerobic Digesters (ADs) (Table 1) grown in isolation or as a mixture in laboratory scale ADs. Both the individual communities and mixes were grown in four replicates. To remove any potentially confounding effects caused by differences in starting density of tested communities, we standardized microbial density based on qPCR-estimated counts of 16S rRNA gene copies. We found that the methane production of the mixed community was initially intermediate between the two individual communities, but after 5 weeks propagation started to produce gas at a rate indistinguishable from the more productive of the individual communities (Figure 1A). We examined both the starting point and the endpoint composition of the single and mixed communities by Illumina sequencing 16s rRNA gene amplicon libraries. Consistent with the phenotypic data, the composition of the mixture was much more similar to the better than the worse performing community at the endpoint (Figure 1B). This was despite the single endpoint communities changing considerably from their ancestral composition over the 5 weeks (Figure 1B).

We next determined if a single community dominated when multiple communities were mixed.

To this end, we propagated 10 single communities (from either industrial ADs or sewage or agricultural waste AD feedstocks, with each replicated three times), and ten replicates of a mixture of all ten communities (Table 1). The results were consistent with those from the two-community mixture. First, methane production in mixtures of ten communities was higher than the average of the individual communities. However, methane production of the mixtures did not differ from the best performing single community, P13, (Figure 2A), which, like each of the single communities used, was a constituent of all the mixtures. Second, the community composition of mixtures (which varied very little between replicates, presumably because they all had the same 10 community starting inocula) most closely resembled the best performing community, P13 (Figure 2B). More generally, the more compositionally similar an individual community was to the replicated 10-community mixtures, the greater the gas production of the community when grown in isolation (Figure S1). Other community characteristics that positively correlated with methane production were bacterial cell densities and within-community (alpha) diversity, but not methanogen density (Figure S2). In summary, the results demonstrate that the community most efficient at using resources (which in these experiments was also the most diverse) dominates when multiple communities are mixed together, thus enhancing mixed community productivity beyond the average of the component communities.

We next explored the ecological mechanism(s) underpinning the observed dominance by the community that produced the most methane. One explanation is that multiple taxa from the same community act as semi-cohesive units and are selected together. This might arise as a result of coevolved mutualistic (or unidirectional) cross-feeding interactions, notably between methanogenic Archaea and hydrogen/acetate producers, where each organism both provides essential resources and removes damaging waste products for each other [12,15,16]. Moreover, coevolved resource partitioning can result in taxa being selected together, because species are expected to coevolve to minimise competition with co-occurring taxa [17–19]. Note that the selection of multiple taxa together in these contexts does not require any form of group selection [11, 20], but simply selection of particular individuals from a key taxon whose presence provides an advantage for individuals from taxa they have coevolved with. This

process can be described as ecological co-selection, equivalent to genetic co-selection, where a gene can hitchhike to high frequency purely as consequence of being linked to genes under positive selection [21].

An alternative explanation is that coevolved interactions within individual communities are relatively unimportant, and the dominant community simply contains more competitive taxa (for any functional task/interaction) than other communities. This does not imply that coevolved cross-feeding interactions are unimportant for methanogenic communities, but that these co-evolved interactions are no more specific for taxa isolated from within a community than taxa isolated from different communities. In other words, functionally equivalent taxa are interchangeable between communities. These different scenarios, selection for the best individual taxa and co-selection, are two extremes of a continuum. The distinction is important because dominance by a single community is necessarily a more likely consequence of community coalescence when co-selection operates. Figure 3 (ABC) provides an illustration of the two extreme scenarios, no co-selection and co-selection of the entire community, and an intermediate case where there are two groups of interacting taxa, or modules, and co-selection occurs within each.

The most direct way to demonstrate a role of co-selection would be to show that the outcome of competition between single taxa from different communities does not predict the outcome of competition at the community level [11]. Unfortunately, this is not feasible for such complex communities, in which many taxa are very difficult to grow in isolation. However, there are other testable predictions associated with the operation of co-selection or otherwise. If the success of an individual taxon is independent of whether they are in the presence of taxa from the same community (i.e. co-selection does not occur), communities that use resources most efficiently and hence achieve the highest biomass per unit of time (productivity) will contain the highest number of the best-performing taxa. It then follows that there will be a positive relationship between the productivity of a community and the proportion of taxa it contributes to the mixture (Figure 3A). If instead taxa are co-selected as modules, the correlation between individual community contribution and productivity is likely to break down.

This is best illustrated by the extreme scenario whereby all taxa within a mixed community are co-selected from a single community: the mixture will be entirely dominated by a single constituent community, hence the contribution of all other communities will be independent of their individual productivity (i.e. they will contribute null to the mixture's composition, even though they have non-zero productivity individually; Figure 3B). The intermediate scenario, where co-selection occurs within two independent modules can also break down this correlation if one module contributes much more to community productivity than the other (Figure 3C).

To determine if co-selection contributed to our findings, we first estimated the contribution of each community to the 10-community mixtures using a non-negative least squares (NNLS) approach. The community that had the most similar composition to the mixtures (and produced the most methane) contributed an estimated 40% of its taxa to the mixtures, with only two other communities contributing more than 10% of their taxa to the mixtures (Figure 3D). We then correlated the contribution each community made to the mixtures with two measures of community productivity: methane production and cell densities (based on 16s rRNA gene copy number), which themselves were positively correlated (Figure S2A). We found no suggestion of a positive correlation between either measure of productivity and contribution to the community (Figure 3D and E). These results suggest that co-selection of taxa played an important role in dominance by the community that produced the most methane.

That community coalescence results in the most productive individual community dominating the mixed community has direct implications for biotechnological uses of microbial communities. Given that the best performing community in isolation largely determined both the composition and performance of mixtures of communities, methane production should increase with increasing number of communities in a mixture. We therefore inoculated laboratory-scale anaerobic digesters with 1, 2, 3, 4, 6 or 12 communities, ensuring that each of the 12 starting communities was used an equal amount of times at each diversity level ([22]; see Table S1). Cumulative methane production over a five-week period increased with

increasing number of communities used as an inoculum (Figure 2C). The positive correlation between community function and the number of inoculating communities is analogous to the commonly observed finding that community productivity increases with increasing species diversity [23]. In this case, the mechanism underlying this positive relationship between the number of communities and productivity is a “sampling effect”: inoculating more communities increases the chance that the best performing community will be present in the mix [24]. However, given that domination of mixtures by one community was not complete (Figure 3D and E), it is possible that mixing communities could increase performance beyond that of the maximum of single communities in some circumstances (transgressive over-yielding, [25]).

Here, we have shown that coalescence of microbial communities results in dominance of a single community, the identity of which can be predicted from its efficiency of resource use in isolation. This dominance is likely to be driven in part by co-selection of interacting taxa within coevolved communities, which is likely to greatly increase the magnitude of dominance following mixing [11]. It is unclear whether such effects would be apparent for aerobic communities where cross-feeding interactions are less important for efficient resource use [26], although studies to date [4] suggest asymmetric outcomes, although less extreme, may be common. Our study has also identified a simple method to significantly improve methane yield during anaerobic digestion: inoculate digesters with a broad range of microbial communities, and the best performing community will dominate. However, further work under a range of conditions is clearly required to determine the generality of these findings. Given that resource use efficiency is often a desirable property of microbial communities, this approach could be applied to a range of biotechnological processes driven by microbial communities, as well as to manipulate microbiomes in clinical and agricultural contexts [13].

AUTHOR CONTRIBUTIONS

Methodology: PS, KM, FB, OSS, PJH, JH, AB; Formal analysis PS, SB, MA and AB; Investigation: PS and FB; Writing - Original Draft: PS and AB; Writing - Review and Edition: All authors; Visualisation PS and AB; Supervision: AB

ACKNOWLEDGEMENTS

The work was funded by the BBSRC, the Royal Society, AXA Research Fund and NERC. KM & JH were funded through the project ENIGME from the INRA metaprogramme MEM (Meta-omics and microbial ecosystems). KM was additionally funded through an Institut Carnot 3BCAR international travel grant.

REFERENCES

1. Rillig, M.C., Antonovics, J., Caruso, T., Lehmann, A., Powell, J.R., Veresoglou, S.D., and Verbruggen, E. (2015). Interchange of entire communities: Microbial community coalescence. Trends Ecol. Evol. *30*, 470–476.
2. Rillig, M.C., Lehmann, A., Aguilar-Trigueros, C.A., Antonovics, J., Caruso, T., Hempel, S., Lehmann, J., Valyi, K., Verbruggen, E., Veresoglou, S.D., *et al.* (2016). Soil microbes and community coalescence. Pedobiologia (Jena). *59*, 37–40.
3. Hausmann, N., and Hawkes, C. (2009). Plant neighborhood control of arbuscular mycorrhizal community composition. New Phytol. *183*, 1188–1200.
4. Livingston, G., Jiang, Y., Fox, J., and Leibold, M. (2013). The dynamics of community assembly under sudden mixing in experimental microcosms. Ecology *94*, 2898–2906.
5. Souffreau, C., Pecceu, B., Denis, C., Rummens, K., and De Meester, L. (2014). An experimental analysis of species sorting and mass effects in freshwater bacterioplankton. Freshw. Biol. *59*, 2081–2095.
6. Adams, H.E., Crump, B.C., and Kling, G.W. (2014). Metacommunity dynamics of bacteria in an arctic lake: The impact of species sorting and mass effects on bacterial production and biogeography. Front. Microbiol. *5*.
7. Calderón, K., Spor, A., Breuil, M.-C., Bru, D., Bizouard, F., Violle, C., Barnard, R.L., and Philippot, L. (2017). Effectiveness of ecological rescue for altered soil microbial communities and functions. ISME J. *11*, 272–283.
8. Gilpin, M. (1994). Community-level competition: asymmetrical dominance. Proc. Natl. Acad. Sci. U. S. A. *91*, 3252–3254.
9. Toquenaga, Y. (1997). Historicity of a simple competition model. J. Theor. Biol. *187*,

175–181.

10. Wright, C.K. (2008). Ecological community integration increases with added trophic complexity. Ecol. Complex. *5*, 140–145.

11. Tikhonov, M. (2016). Community-level cohesion without cooperation. Elife *5*, e15747.

12. Schink, B. (1997). Energetics of syntrophic cooperation in methanogenic degradation. Microbiol. Mol. Biol. Rev. *61*, 262–280.

13. Rillig, M.C., Tsang, A., and Roy, J. (2016). Microbial community coalescence for microbiome engineering. Front. Microbiol. *7*, 6–8.

14. Hillesland, K.L., and Stahl, D. a (2010). Rapid evolution of stability and productivity at the origin of a microbial mutualism. Proc. Natl. Acad. Sci. U. S. A. *107*, 2124–2129.

15. Embree, M., Liu, J.K., Al-Bassam, M.M., and Zengler, K. (2015). Networks of energetic and metabolic interactions define dynamics in microbial communities. Proc. Natl. Acad. Sci. U. S. A. *112*, 15450–15455.

16. Großkopf, T., and Soyer, O. (2016). Microbial diversity arising from thermodynamic constraints. ISME J. *10*, 2725–2733.

17. Schluter, D. (2000). The ecology of adaptive radiation.

18. Roughgarden, J. (1976). Resource partitioning among competing species - A coevolutionary approach. Theor. Popul. Biol. *9*, 388–424.

19. MacArthur, R.H. (1970). Species-packing and competitive equilibrium for many species. Theorectical Popul. Biol. *1*, 1–11.

20. Gardner, A., and Grafen, A. (2009). Capturing the superorganism: a formal theory of group adaptation. J. Evol. Biol. *22*, 659–671.

21. Baker-Austin, C., Wright, M.S., Stepanauskas, R., and McArthur, J. V. (2006). Co-selection of antibiotic and metal resistance. Trends Microbiol. *14*, 176–182.

22. Hodgson, D.J.D., Rainey, P.B., and Buckling, A. (2002). Mechanisms linking diversity, productivity and invasibility in experimental bacterial communities. *269*, 2277–2283.

23. Tilman, D., and Lehman, C. (1997). Plant diversity and ecosystem productivity: theoretical considerations. Proc. Natl. Acad. Sci. U. S. A. *94*, 1857–1861.

24. Tilman, D. (1999). The ecological consequences of changes in biodiversity: A search for general principles. In Ecology (Ecological Society of America), pp. 1455–1474.

25. Harper, D. (1977). The population biology of plants (Academic Press).

26. Morris, B.E.L., Henneberger, R., Huber, H., and Moissl-Eichinger, C. (2013). Microbial syntrophy: Interaction for the common good. FEMS Microbiol. Rev. *37*, 384–406.

27. Einen, J., Thorseth, I.H., and Øvreås, L. (2008). Enumeration of Archaea and Bacteria in seafloor basalt using real-time quantitative PCR and fluorescence microscopy. FEMS Microbiol. Lett. *282*, 182–187.

28. Ruijter, J.M., Ramakers, C., Hoogaars, W.M.H., Karlen, Y., Bakker, O., van den hoff, M.J.B., and Moorman, A.F.M. (2009). Amplification efficiency: Linking baseline and bias in the analysis of quantitative PCR data. Nucleic Acids Res. *37*, e45.

29. Brankatschk, R., Fischer, T., Veste, M., and Zeyer, J. (2013). Succession of N cycling processes in biological soil crusts on a Central European inland dune. FEMS Microbiol. Ecol. *83*, 149–160.

30. Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., and Schloss, P.D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl. Environ. Microbiol. *79*, 5112–5120.

31. Eren, A.M., Maignien, L., Sul, W.J., Murphy, L.G., Grim, S.L., Morrison, H.G., and Sogin, M.L. (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol. Evol. *4*, 1111–1119.

32. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods *7*, 335–336.

33. Edgar, R.R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

34. Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. Bioinformatics *27*, 2194–2200.

35. McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. *6*, 610–8.

36. Mullen, K., and Van Stokkum, I. (2007). The Lawson-Hanson algorithm for non-negative least squares (NNLS).

37. Soetaert, K., Meersche, K.V.D., and Oevelen, D. V (2009). Package limSolve, solving linear inverse models in R.

FIGURE LEGENDS

Figure 1: Temporal dynamics of methane production and composition when two communities are mixed. A) Cumulative methane production in ml (±SEM) over time of: community P01 (white circles), community P05 (black circles) and their mixes (grey circles). Cumulative methane production differed between treatments (ANOVA: $F_{2,9}$ = 23.2, P < 0.001), but did not differ between the mixed community and P05 (Tukey-Kramer HSD: P = 0.5). P01 was lower than both other treatments (Tukey-Kramer HSD: P < 0.001 in both cases). B) NMDS plot of unweighted UniFrac of communities P01 (white), P05 (black) and their mixes (grey). Ancestral samples are represented by squares with samples from the endpoint of the experiment by circles. At the endpoint, P05 was compositionally more similar to the mixtures than P01, based on both unweighted (*t-tests* of mean distance to each mixture for each replicate single community: t_6 = 8.3, P < 0.001) and weighted (t_6 = 2.3, P = 0.03) UniFrac distances.

Figure 2: Methane production and community composition when multiple communities are mixed. A) Total methane production of mixed (grey) and individual communities (white), with mean values shown as horizontal lines. Mean total methane production was greater for mixtures than for individual communities (*t*-test: P < 0.001 in 9 cases), except when measured against community P13 (the best performer). B) NMDS plot of unweighted unifrac of 10 mixtures (grey) and 9 individual communities (white). Numbers in circles refer to individual community identifiers (Table 1). Community P13 was significantly closer in composition to the 10 mixed communities than any other community (weighted and unweighted UniFrac distances; Paired *t*-tests; P < 0.001, in all cases). There was also a significant link between the community composition and the difference in gas production between the communities (see Figure S1). Note: DNA yield from community P06, which had the lowest gas production

of all communities, was insufficient for sequencing, therefore it is excluded from this and following graphs. C) Individual communities (white circles) and their average methane production (white line); mixes of communities (grey circles) and their averages (grey line). There was a monotonic increase in methane production with number of communities used (Regression: $F_{1,26} = 5.4$, $P = 0.03$). For community composition of the mixes, see Table 1 and Table S1.

Figure 3: The role of co-selection in explaining dominance by a single community. A-C) The top panels illustrate three hypothetical scenarios describing how communities contribute to a mixture of communities, while the bottom panels show the expected relationships between a community's contribution and its methane production. The letters within the top panels indicate taxa that drive two biochemical processes (abcd and ef); capitalised letters are the best representatives of a taxon among all the communities. A). No co-selection. B). Co-selection of all taxa within a community. C). Co-selection of taxa within two independent modules. D). Mean estimated relative contribution of each individual community (numbered) towards the 10 coalesced communities calculated using the NNLS method, plotted against mean cumulative methane production for each community; there is no significant relationship (Regression; $F_{1,7} = 1.7$, $P > 0.2$). E) As D, but relative contribution is plotted against number of bacterial and archaeal cells calculated based on the 16S rRNA gene copy number (Regression; $F_{1,7} = 1.7$, $P > 0.5$. Note the relative contribution is not a fractional contribution because some OTUs present in the mixture were not detected in the constituent communities. This is presumably because they only reached detected frequencies in the mixture, but we can't rule out that the community that we failed to get sufficient reads from contributed to the composition of the mixtures. Mind that the cell densities of Archaea and bacteria do not significantly correlate with the gas production (see Figure S2).

Table 1: **List of individual communities used in this analysis and their source.** All Anaerobic Digester (AD) communities were derived from industrial ADs in the South West of

England. Specific locations cannot be provided because of commercial sensitivity. Note that experiment numbers correspond with figure numbers.

Sample name	Feed/Type	Temperature	Used in experiments
P01	Silage and Foodwaste Anaerobic Digester (AD)	44 - 42.5°C	1,2,3
P02	Silage + Food waste AD	44 - 42.5°C	2,3
P03	Maize/Cow Slurry/Chicken Manure AD	45°C	3
P04	Maize/Cow Slurry/Chicken Manure AD	45°C	2,3
P05	Sewage Sludge AD	36°C	1,2,3
P06	Raw Sewage	Ambient	2,3
P08	Thickened Sewage Sludge	Ambient	2,3
P09	Sewage Based AD	36°C	2,3
P10	Food Waste AD	36°C	2,3
P11	Cow Slurry	Ambient	3
P12	Silage, Slurry and Manure Pre-Digestate	Ambient	3
P13	Silage, Slurry and Manure AD	40°C	2,3
P15	Food waste AD	36°C	2

STAR METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

The authors are happy to share any further resources linked to the research involved with qualified third parties. Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Pawel Sierocinski (p.sierocinski@exeter.ac.uk).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Methanogenic communities

The communities used were collected from anaerobic digesters (AD plants: communities P01,P02, P03, P04, P05, P09, P10, P13 and P15) and communities present in nature used to seed the AD plants (communities P06, P08, P11, P12, see the details in Table 1). All the communities have been collected in the South West area of United Kingdom from operating Anaerobic Digesters and the seeding communities they use for the reactors. The

communities were operating at temperatures between ambient and 45°C in their places of origin. Communities were stored at 4°C prior to use.

METHOD DETAILS

Cultivation details

For all experiments, communities were grown in 500 ml bottles (600ml total volume with headspace; Duran) using the commercially available Automated Methane Potential Test System (AMPTS, Bioprocess Control Sweden AB) to measure CO_2-stripped biogas production (referred to as methane in this paper). Samples were fed weekly with 25 ml of defined medium in a fed-batch mode using a defined medium (see below for media composition).

The communities used in experiment 1 were equalised in terms of bacterial cells per gram of sample before inoculation using M9 salts to dilute them to the community with the lowest cell density, based on qPCR enumeration of 16S rRNA gene copies. For experiments 2 and 3, starting 16S rRNA copy number was determined (but not equalised between communities) and did not correlate with methane production. The fermenters were inoculated with 275 g of sample and fed with 25 ml of defined medium: meat extract 111.1 gl^{-1}, cellulose 24.9 gl^{-1}, starch 9.8 gl^{-1} glucose 0.89 gl^{-1}, xylose 3.55 gl^{-1} (carbon to nitrogen ratio of 15:1) every week, starting with t_0. Before the start of the fermentation, 0.3 mL of 1000x Trace Metal stock (1 gl^{-1} $FeCl_2 \cdot 4H_2O$, 0.5 gl^{-1} $MnCl_2 \cdot 4H_2O$, 0.3 gl^{-1} $CoCl_2 \cdot 4H_2O$, 0.2 gl^{-1} $ZnCl_2$, 0.1 gl^{-1} $NiSO_4 \cdot 6H_2O$, 0.05 gl^{-1} $Na_2MoO_4 \cdot 4H_2O$, 0.02 gl^{-1} H_3BO_3, 0.008 gl^{-1} $Na_2WO_4 \cdot 2H_2O$, 0.006 gl^{-1} $Na_2SeO_3 \cdot 5H_2O$, 0.002 gl^{-1} $CuCl_2 \cdot 2H_2O$) was added to each fermenter.

Experiment structure

In Experiment 1 (results shown in Figure 1) we cultivated community P01 (four replicates) and community P05 (four replicates) and a 1:1 mix of the two. It ran for 5 weeks before samples were harvested for sequencing (see below). The initial community was sequenced at the same time. In Experiment 2 (results shown in Figure 2), we cultivated 10 individual

communities (listed in Table 1), in triplicate, and 10 mixes of all 10 communities mixed in equal volumes, at the same total volume as the single communities. After 6 weeks samples were harvested for sequencing. In Experiment 3 (results shown in Figure 2C) we used 12 communities (detailed in Table 1). They were grown in isolation as well as pseudo-randomly combined to create a gradient of number of starting communities, with each community used only once for each number of communities. This resulted in 12 single communities, 6 pairs, 4 triplets, 3 quadruplets, 2 mixes of 6 and one mix of 12 communities. Specific details of mixing can be seen in Table S1. The cultures were propagated for 5 weeks.

Measuring methane content of Biogas

All resulting lab-scale reactors inoculated with the samples were run at 37°C using the Automatic Methane Potential Test System (AMPTS). The AMPTS is a setup of 15 simple fermenters using a 0.5L lab bottle as the vessel with its own stirring system provided with a butyl rubber stopper and sampling ports. It is connected to an online gas measuring system to allow continuous gas measurements. The AMPTS system measures the volume of biogas produced following stripping of CO_2 (by passing the gas through 50 ml of 3M NaOH solution) from the produced gas. To reproduce our results, however, there is no need for a sophisticated setup, some pilot experiments yielding similar results in terms of gas production were conducted using anaerobic serum flasks. We confirmed that the measured biogas was >95% methane using Gas Chromatography with Flame Ionisation Detection optimized for methane detection.

DNA extraction and qPCR quantification

DNA for 16S rRNA gene amplicon sequencing was extracted using QIAamp DNA Stool Mini Kit (QIAGEN) or FastDNA™ SPIN Kit for Soil (MP), depending on the experiment. Note that DNA extraction for mixed community P06 from experiment 2 failed. The DNA for qPCR was extracted with the QIAamp DNA Stool Mini Kit (QIAGEN), protocol for pathogen detection with the 95ºC incubation step and the Powerlyzer Powersoil DNA KIT (MOBIO). DNA from *Acinetobacter baylyi*, *Pseudomonas fluorescens* SBW25 for Bacteria and from *Halobacterium salinarum* DSM 669 for Archaea was used as standards. The primers [27] used to quantify

Bacteria were 16S rRNA 338f - ACT CCT ACG GGA GGC AGC AG, 518r - ATT ACC GCG GCT GCT GG for Archaea: 931f - AGG AAT TGG CGG GGG AGC A, m1100r - BGG GTC TCG CTC GTT RCC. The reagents used were: 1x Brilliant III Ultra-Fast SYBR® Green QPCR Master Mix; 150nM 338f and 300nM 518r or 300nM 931f and 300 nM m1100r; ROX 300nM; and BSA 100 ng/µl final concentration. All samples were run in triplicate on a StepOnePlus (Applied Biosystems) qPCR machine using a program with 3 minutes 95ºC initial denaturation followed by 40 cycles of 5 seconds at 95ºC and 10 seconds at 60ºC, followed by a melting curve 95ºC for 15 seconds; 60ºC for 1 minute ramping up to 95ºC in steps of 0.3ºC for 15 seconds each. The melting curve analysis and the confirmation of the negative controls was done using StepOne Software v.2.3 (life technologies). The Cq values and the efficiencies of the samples and standards was determined as previously using LinRegPCR version 2016.0[28]. The quantities were calculated using the one point calibration method as described earlier[29].

Amplicon library construction and sequencing

16S rRNA gene libraries were constructed using primers designed to amplify the V4 region and multiplexed [30]. Amplicons were generated using a high-fidelity polymerase (Kapa 2G Robust) and purified using the Agencourt AMPure XP PCR purification system and quantified using a fluorometer (Qubit, life technologies). The purified amplicons were then pooled in equimolar concentrations by hand based on Qubit quantification. The resulting amplicon library pool was diluted to 2 nM with sodium hydroxide and 5 µl transferred into 995 µl HT1 (Illumina) to give a final concentration of 10 pM. 600 µl of the diluted library pool was spiked with 10% PhiX Control v3 and placed on ice before loading into Illumina MiSeq cartridge following the manufacturer's instructions. The sequencing chemistry utilised was MiSeq Reagent Kit v2 (500 cycles) with run metrics of 250 cycles for each paired end read using MiSeq Control Software 2.2.0 and RTA 1.17.28.

Analyses of sequenced samples

MiSeq amplicon reads were merged using Illumina-utils software [31]. We quality-filtered only the mismatches in the overlapping region between read pairs using a minimum overlap (--

min-overlap-size) of 200 nt and a minimum quality Phred score (--min-qual-score) of Q20. We allowed no more than five mismatches per 100 nt (-P 0.05) over the 200 nt overlapping region.

Reads that fulfilled the quality criteria were analysed using Quantitative Insights Into Microbial Ecology (QIIME v.1.7) [32]. We removed chimera using the *identify_chimeric_seqs.py* script, UCHIME reference 'Gold' database and USEARCH [33,34], which we also used to select OTUs. We assigned the taxonomy of our reads with QIIME *pick_open_reference_otus.py* function, using the Greengenes database version v13_8[35] as a reference with a minimum cluster size of 2 (i.e., each OTU must contain at least two sequences). We collapsed the technical replicates and filtered out the low abundance OTUs (<0.01% total, *filter_otus_from_otu_table.py*) and samples rarefied to an even depth of 26702 for both experiments where sequencing data is available. QIIME was used to calculate alpha and beta diversity data and produce NMDS plots.

QUANTIFICATION AND STATISTICAL ANALYSIS

MacQiime was used to calculate alpha and beta diversity data and produce NMDS plots. Data obtained with MacQiime was later combined with the gas production data and analysed using JMP Pro 13 software (SAP) as described in the Figure legends.

For the NNLS analysis, following removal of low abundance OTUs and cumulative sum scaling transformation, the resulting biom file was used to create a matrix $A \in \mathbb{Z}_{\geq 0}^{m \times n}$ (*m* rows of OTUs by *n* sample columns) for all of the single communities, and a column vector $b \in \mathbb{Z}_{\geq 0}^{m}$ for each of the mixed communities; both A and b hold non-negative integers of OTU abundances. Note that one of the individual samples contained a negligible number of reads and was discarded from the analysis. The contribution, or weight, of each seed sample to the pattern of OTUs observed in each of the mixed communities is given by the column vector $x \in \mathbb{R}^n$ when solving a system of linear equations $Ax = b$. Written out this equation ($Ax = b$) looks like this for each mixture:

$$\begin{pmatrix} \mathrm{OTU}_{1,S1} & \cdots & \mathrm{OTU}_{1,Sn} \\ \vdots & \ddots & \vdots \\ \mathrm{OTU}_{m,S1} & \cdots & \mathrm{OTU}_{m,Sn} \end{pmatrix} \mathrm{x} \begin{pmatrix} x_{S1} \\ \vdots \\ x_{Sm} \end{pmatrix} = \begin{pmatrix} OTU_{mix\_1} \\ \vdots \\ OTU_{mix\_m} \end{pmatrix}$$

_____

where S refers to each single community.

When modelling count data for environmental samples the fitted parameters of x will also be non-negative and the number of OTUs will usually exceed the number of samples (*m* > *n*). The task is to solve an over-determined system of linear equations where there are more equations than unknowns. It is likely that some of the linear equations will 'disagree' and there will be no exact solution. Geometrically, this may be interpreted as b not lying in the column space of A, a (hyper)plane holding the column vectors of A, or $Ax - b \neq 0$. A least-squares approach may find the non-negative vector $\bar{x} = (A^T A)^{-1} A^T b$ which is the projection of b back onto the column space of A that minimises the least-squares error 'distance' $\|A\bar{x} - b\|$. For our study the non-negative least-squares (NNLS) solution, $\bar{x}$, and least-squares errors were computed via the R packages 'nnls' [36] and 'limSolve' [37] for each of the mixed samples, with the mean relative contribution of each community to the 10 mixtures shown in Figure 2C (relative to the gas production)..

DATA AND SOFTWARE AVAILABILITY

The raw sequences obtained from our experiments are available at the European Nucleotide Archive and may be accessed at http://www.ebi.ac.uk/ena/data/view/PRJEB21193 (Experiment 1) and http://www.ebi.ac.uk/ena/data/view/PRJEB21187 (Experiment 2). We also included the R code that allows the user to calculate the contribution a single community has in a mix of communities (see Method S1). Using this code, a NNLS analysis can be conducted with the input of a pre-filtered OTU table.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Bacterial and Virus Strains		
Biological Samples		
Chemicals, Peptides, and Recombinant Proteins		
Brilliant III Ultra-Fast SYBR (R) Green QPCR Master Mix	Agilent Technologies	600882-51
meat extract	Sigma-Aldrich Co. LLC.	70164-500G
Xylose	Sigma-Aldrich Co. LLC.	W360100-1KG
Cellulose	Sigma-Aldrich Co. LLC.	C6288
Starch	Sigma-Aldrich Co. LLC.	33615-1KG
Glucose	Sigma-Aldrich Co. LLC.	G8270-1KG
Critical Commercial Assays		
QIAamp DNA Stool Mini Kit (QIAGEN)	Qiagen	ID: 51504
FastDNA™ SPIN Kit for Soil	MP Biomedicals, LLC	116560200
PowerLyzer® PowerSoil® DNA Isolation Kit	MO BIO Laboratories, Inc.	12855-100
Deposited Data		
Sequencing data Experiment 1	European Nucleotide Archive	http://www.ebi.ac.uk/ena/data/view/PRJEB21193

Sequencing data Experiment 2	European Nucleotide Archive	http://www.ebi.ac.uk/ena/data/view/PRJEB21187
Experimental Models: Cell Lines		
Experimental Models: Organisms/Strains		
Community P01	Silage and Foodwaste Anaerobic Digester (AD)	This paper
Community P02	Silage + Food waste AD	This paper
Community P03	Maize/Cow Slurry/Chicken Manure AD	This paper
Community P04	Maize/Cow Slurry/Chicken Manure AD	This paper
Community P05	Sewage Sludge AD	This paper
Community P06	Raw Sewage	This paper
Community P08	Thickened Sewage Sludge	This paper
Community P09	Sewage Based AD	This paper
Community P10	Food Waste AD	This paper
Community P11	Cow Slurry	This paper
Community P12	Silage, Slurry and Manure Pre-Digestate	This paper
Community P13	Silage, Slurry and Manure AD	This paper
Community P15	Food waste AD	This paper
Oligonucleotides		
338f - ACT CCT ACG GGA GGC AGC AG	[27]	
518r - ATT ACC GCG GCT GCT GG	[27]	
931f - AGG AAT TGG CGG GGG AGC A	[27]	
m1100r - BGG GTC TCG CTC GTT RCC	[27]	
Recombinant DNA		
Software and Algorithms		

StepOne Software v.2.3	life technologies	https://www.thermofisher.com/uk/en/home/technical-resources/software-downloads/StepOne-and-StepOnePlus-Real-Time-PCR-System.html#
LinRegPCR version 2016.0	[28]	linregpcr.nl
R version 3.4.0	R Core Team (2013).	Mac: https://cran.r-project.org/bin/macosx/ PC: https://cran.r-project.org/bin/windows/base/old/
macQIIME	[32]	http://www.wernerlab.org/software/macqiime/download
Other		
NNLS Method for assessing community contribution in a mix	This paper,	Method S1

Figure 1

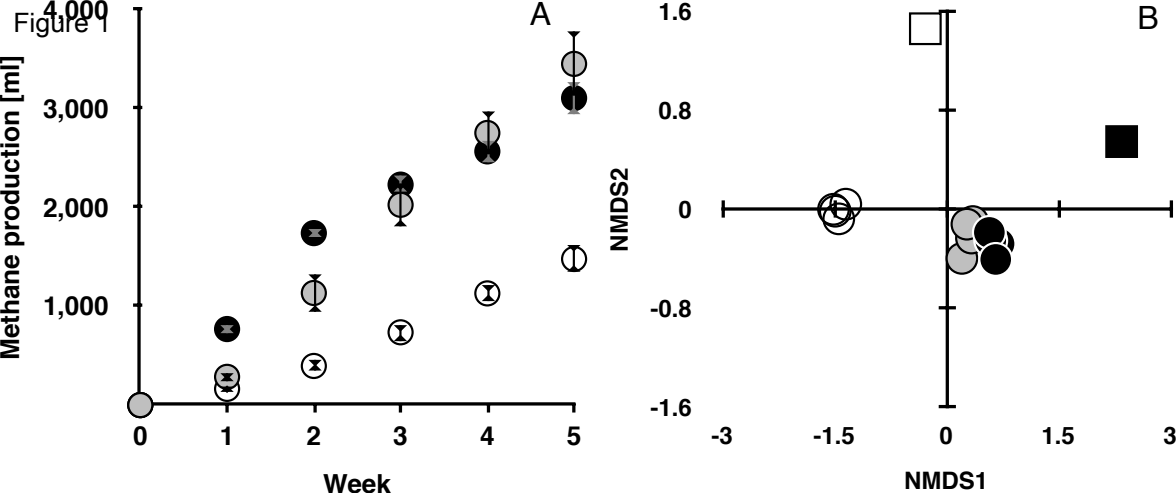

A
Methane production [ml]
4,000
3,000
2,000
1,000
0
1
2
3
4
5
Week
B
NMDS2
1.6
0.8
0
-0.8
-1.6
-3
-1.5
0
1.5
3
NMDS1

Figure 2

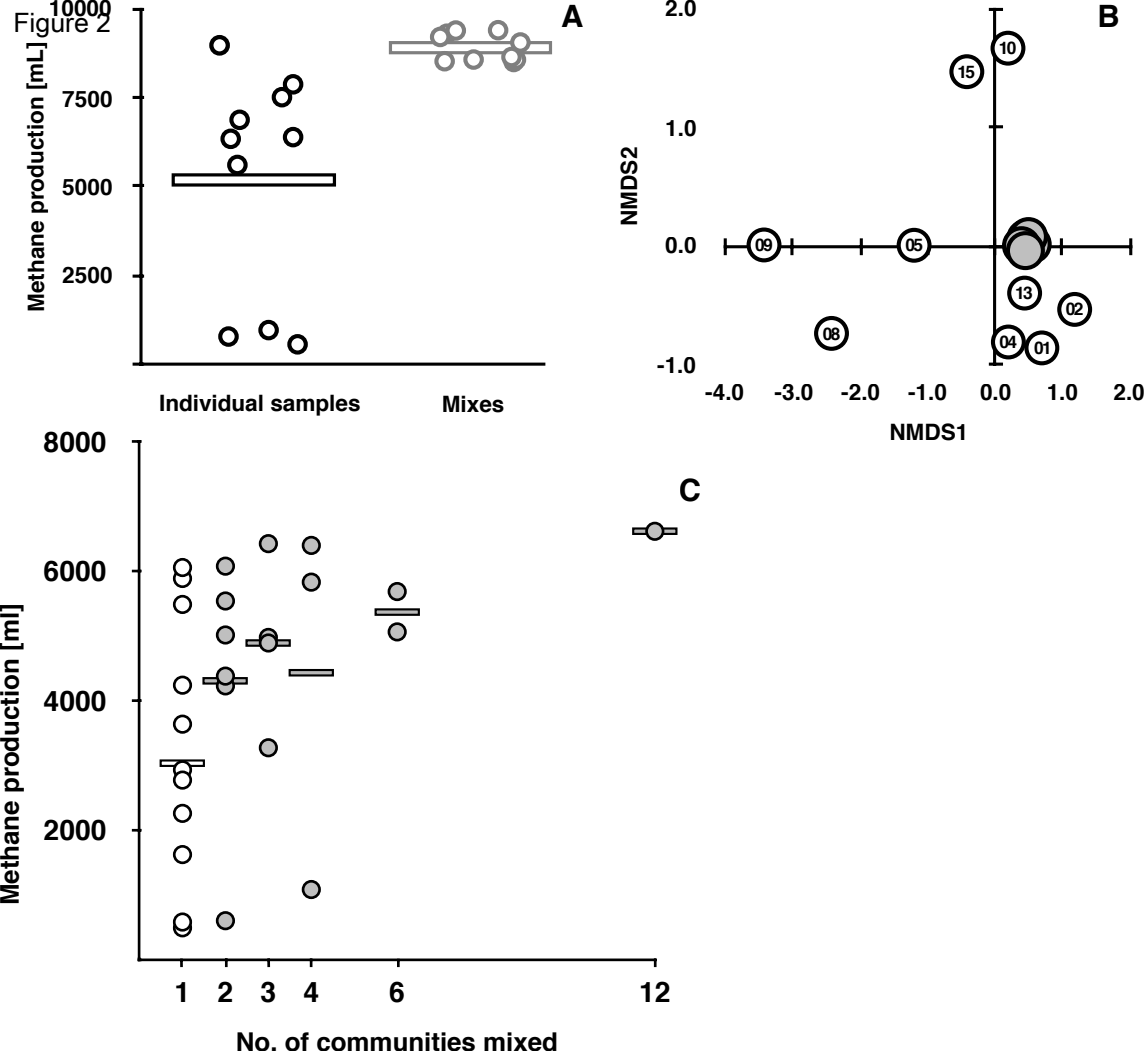
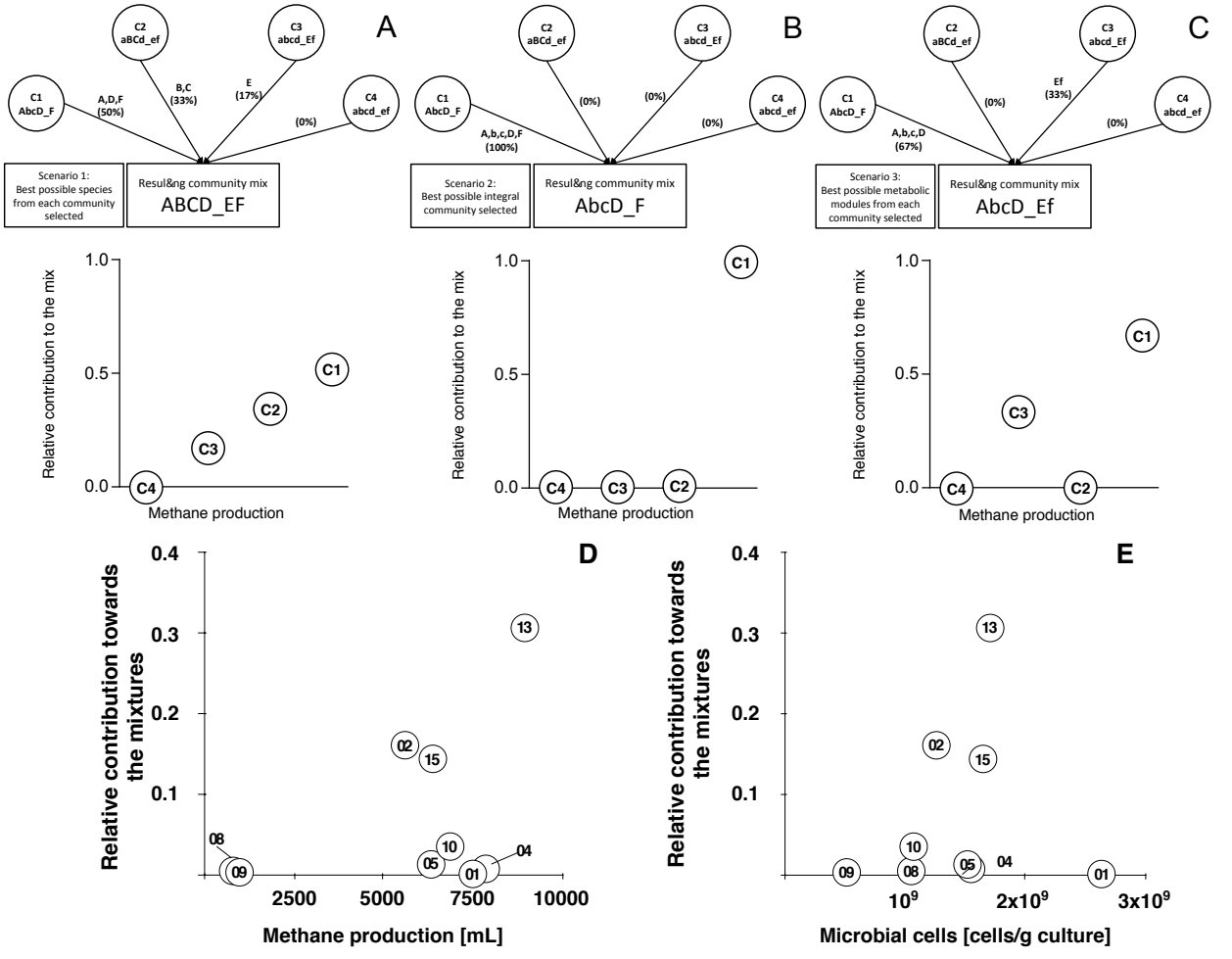

A
Methane production [mL]
10000
7500
5000
2500
Individual samples
Mixes
B
NMDS2
2.0
1.0
0.0
-1.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
NMDS1
10
15
09
05
13
02
08
04
01
C
Methane production [ml]
8000
6000
4000
2000
1
2
3
4
6
12
No. of communities mixed

Figure 3


A
C1 AbcD_F
C2 aBCd_ef
C3 abcd_Ef
C4 abcd_ef
A,D,F (50%)
B,C (33%)
E (17%)
(0%)
Scenario 1: Best possible species from each community selected
Resul&ng community mix
ABCD_EF
B
A,b,c,D,F (100%)
Scenario 2: Best possible integral community selected
AbcD_F
C
A,b,c,D (67%)
Ef (33%)
Scenario 3: Best possible metabolic modules from each community selected
AbcD_Ef
Relative contribution to the mix
Methane production
D
Relative contribution towards the mixtures
Methane production [mL]
E
Microbial cells [cells/g culture]

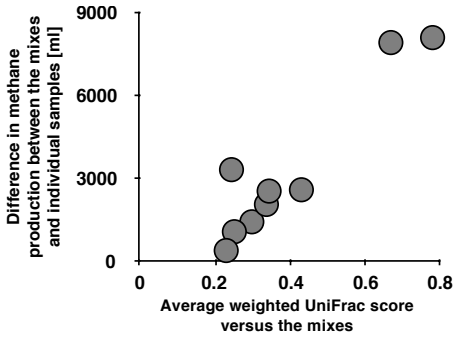


Figure S1: The relationship between the difference in composition of single communities from the mixtures and the difference in gas production of single communities from the mixture related to Figures 2A, 2B.. The Y-axis shows the difference in gas production between the mean of three replicates of each individual community and the mean gas production of the mixes (the smaller the value, the more similar the gas production to the mixtures). The X-axis shows the mean unweighted UniFrac distance between individual communities and each of the ten mixtures (the smaller the value, the more similar the community composition to the mixtures). These two variables are positively correlated (Spearman $\rho = 0.86$, $P < 0.001$)). Qualitatively similar results were obtained using weighted UniFrac distances (Spearman $\rho = 0.75$, $P < 0.02$).

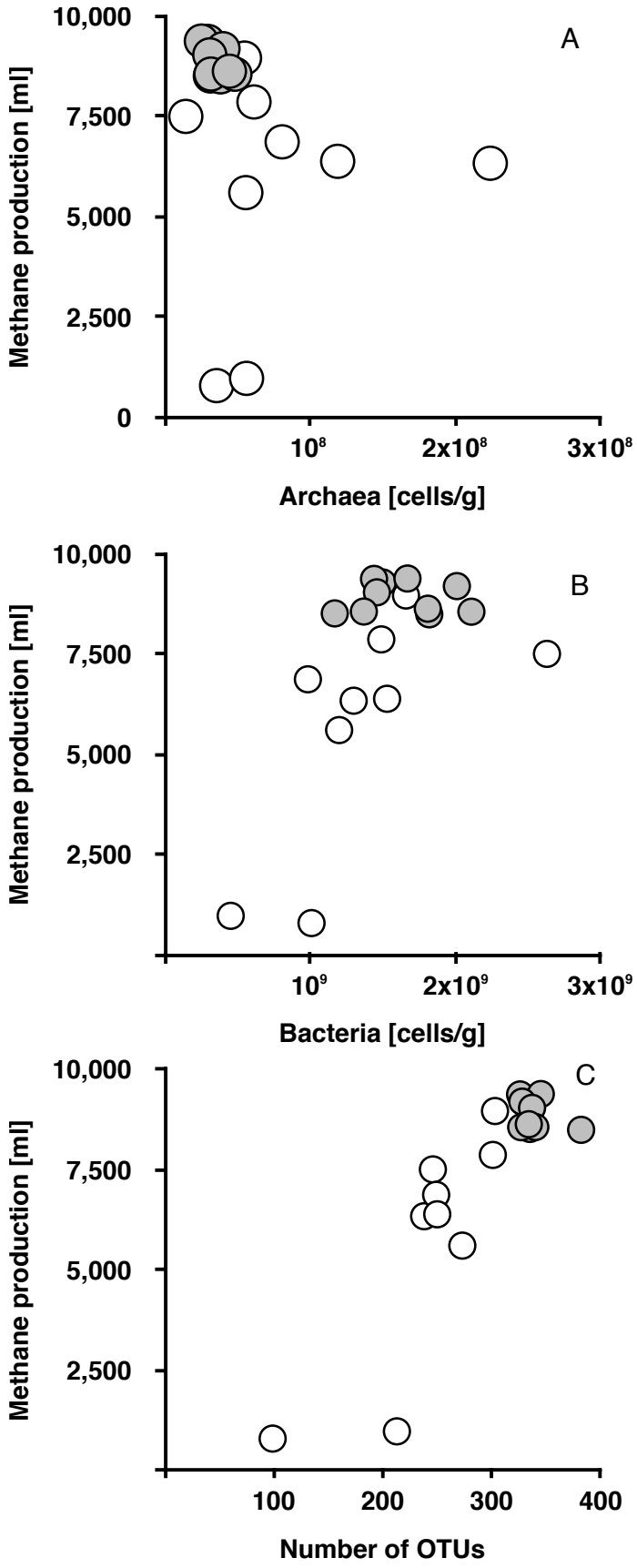

A
Methane production [ml]
10,000
7,500
5,000
2,500
0
10^8
2x10^8
3x10^8
Archaea [cells/g]
B
Methane production [ml]
10,000
7,500
5,000
2,500
10^9
2x10^9
3x10^9
Bacteria [cells/g]
C
Methane production [ml]
10,000
7,500
5,000
2,500
100
200
300
400
Number of OTUs

Figure S2: Within-community predictors of methane production from Experiment 2 related to Figure 3D, 3E. Relationships between methane production [ml] and: A) Archaeal densities [$\log_{10}$ cells/g] (Regression: $F_{1,15} = 0.32$, $P > 0.2$); B) Bacterial densities [$\log_{10}$ cells/g] (Regression: $F_{1,16} = 16.5$, $P < 0.001$); and C) number of OTUs (Regression: $F_{1,16} = 51.6$, $P < 0.001$). The reported statistics are based on combined mixed (grey circles) and individual (white circles) communities, but the same qualitative relationships were found when mixed communities were excluded form the analyses (Archaeal density: $F_{1,7} = 0.07$, $P > 0.2$; Bacterial density: $F_{1,7} = 92$, $P < 0.02$; OTUs: $F_{1,7} = 16.4$, $P < 0.01$).

Sample consisting of community(ies)	No. of communities in the mix	Gas production after 4 weeks	Gas production of the best component of the mix after 4 weeks	Average gas production of all components of a mix after 4 weeks	Difference in gas production between a mix and its best component	Difference in gas production between a mix and the average of its components
P06	1	494.9	N/A	N/A	N/A	N/A
P08	1	584.3	N/A	N/A	N/A	N/A
P09	1	600.8	N/A	N/A	N/A	N/A
P12	1	1626.2	N/A	N/A	N/A	N/A
P11	1	2262.8	N/A	N/A	N/A	N/A
P03	1	2776.8	N/A	N/A	N/A	N/A
P01	1	2935.5	N/A	N/A	N/A	N/A
P02	1	3640.8	N/A	N/A	N/A	N/A
P05	1	4243.0	N/A	N/A	N/A	N/A
P10	1	5490.1	N/A	N/A	N/A	N/A
P04	1	5891.3	N/A	N/A	N/A	N/A
P13	1	6060.3	N/A	N/A	N/A	N/A
P01+P11	2	4231.70	2935.50	2599.15	1296.20	1632.55
P02+P09	2	5018.40	3640.80	2120.80	1377.60	2897.60
P12+P03	2	604.70	2776.80	2201.50	-2172.10	-1596.80
P08+P05	2	4383.00	4243.00	2413.65	140.00	1969.35
P04+P06	2	5543.40	5891.30	3193.10	-347.90	2350.30
P13+P10	2	6081.40	6060.30	5151.65	21.10	929.75
P03+P12+P01	3	3276.30	2935.50	2446.17	340.80	830.13
P08+P02+P06	3	4980.60	3640.80	1573.33	1339.80	3407.27
P04+P11+P13	3	6429.10	6060.30	4738.13	368.80	1690.97
P05+P09+P10	3	4894.80	5490.10	3444.63	-595.30	1450.17
P01+P13+P10+P05	4	5834.60	6060.30	4682.23	-225.70	1152.38
P02+P04+P11+P06	4	6401.90	5891.30	3072.45	510.60	3329.45
P03+P08+P09+P12	4	1086.50	2776.80	1397.03	-1690.30	-310.53
P12+P02+P01+P09	6	5065.90	4243.00	2271.77	822.90	2794.13

+P08+P05						
P03+P04+P06+P10 +P11+P13	6	5689.20	6060.30	3829.37	-371.10	1859.83
All 12 communities' mix	12	6620.00	6060.30	3050.57	559.70	3569.43
AVERAGE FOR ALL MIXES		4758.8	4672.9	3011.6	85.9	1747.2

Table S1: Community mixing setup and detailed results of Experiment 3 related to Figure 2C. The details of the communities used can be found in Table 1.

Method S1:

```
#####################################################################
## Solving a system of linear equations for a non-square [m > n]   ##
## matrix via the use of non-negative least squares [NNLS].        ##
## Return the solution 'weights' and residual sum of squares.      ##
##                                                                 ##
## Mark Alston, Earlham Institute: mark.alston@earlham.ac.uk       ##
#####################################################################


## ================================================================================================================ ##
## Install the R packages phyloseq, biomformat, nnls and limSolve.
## Now load as required...
## ================================================================================================================ ##

library("phyloseq")
packageVersion("phyloseq")

library("biomformat")
packageVersion("biomformat")

library("nnls")
packageVersion("nnls")


## ================================================================================================================ ##
## READ IN THE OTU table
## ================================================================================================================ ##
## MAPPING FILE TO BE USED: 'mappingFile.txt'
## OTU TABLE TO BE USED:    'CSS_norm_rawValues.biom'

## SET YOUR WORKING DIRECTORY, e.g. point to where your files are located
setwd("/path/to/the/data")


## Quick look at the OTU table which is in biom format
read_biom("CSS_norm_rawValues.biom")


## load the biom file into phyloseq
data = import_biom("CSS_norm_rawValues.biom")
data


## ================================================================================================================ ##
## Give meaningful names to the Taxon Column Headers and create a phyloseq object
## ================================================================================================================ ##
myTaxTable <- tax_table(data)
colnames(myTaxTable) <- c("Kingdom","Phylum", "Class", "Order", "Family", "Genus","Species")

### take a look ###
head(myTaxTable)
rank_names(myTaxTable)

OTU = otu_table(data)
TAX = tax_table(myTaxTable)

myPhyloSeq_allData <- phyloseq(OTU,TAX)
myPhyloSeq_allData


## ================================================================================================================ ##
## Incorporate some metadata about the samples ### i.e. first create a mappingFile.txt in a text editor
## ================================================================================================================ ##
SAMPLES = import_qiime_sample_data("mappingFile.txt")
class(SAMPLES)



## ================================================================================================================ ##
## MERGE the bits and bobs into a new phyloseq object
## ================================================================================================================ ##
myPhyloSeq <- merge_phyloseq(myPhyloSeq_allData, SAMPLES)
myPhyloSeq


## inspect the OTU table
otu_table(myPhyloSeq)

## Checking the column headers and looking at the mapping file we see that
## columns 3,2,1,6,18,17,4,7,5 correspond to the 9 SINGLE communities [SAM11640-SAM11649]
## columns 11,12,10,19,15,16,13,14,8,9 correspond to the 10 MIXED communities [SAM11630-SAM11639]


## ================================================================================================================ ##
## collapse OTU table at the 'family' level, and generate matrix 'A' [A.X = b]
## ================================================================================================================ ##

bacteria_family <- tax_glom(myPhyloSeq, taxrank="Family")
bacteria_family_df <- as.data.frame(get_taxa(otu_table(bacteria_family))  )
bacteria_family_singleComm_df  <-  bacteria_family_df[,c(3,2,1,6,18,17,4,7,5)]

row.names(bacteria_family_singleComm_df) <- NULL
colnames(bacteria_family_singleComm_df) <- NULL
```

```
## ============================================================================================================ ##
## get matrix 'A'
## ============================================================================================================ ##
bacteria_family_matrix_A <- as.matrix(bacteria_family_singleComm_df)



## ============================================================================================================ ##
## get vector 'b' [A.X = b], one for each of the 'mixed' samples [OTU table columns 11,12,10,19,15,16,13,14,8,9]
## ============================================================================================================ ##
b_M01_bac_family_df  <- bacteria_family_df[,c(11)]
bf_M01  <-  as.matrix(b_M01_bac_family_df)

b_M02_bac_family_df  <- bacteria_family_df[,c(12)]
bf_M02  <-  as.matrix(b_M02_bac_family_df)

b_M03_bac_family_df  <- bacteria_family_df[,c(10)]
bf_M03  <-  as.matrix(b_M03_bac_family_df)

b_M04_bac_family_df  <- bacteria_family_df[,c(19)]
bf_M04  <-  as.matrix(b_M04_bac_family_df)

b_M05_bac_family_df  <- bacteria_family_df[,c(15)]
bf_M05  <-  as.matrix(b_M05_bac_family_df)

b_M06_bac_family_df  <- bacteria_family_df[,c(16)]
bf_M06  <-  as.matrix(b_M06_bac_family_df)

b_M07_bac_family_df  <- bacteria_family_df[,c(13)]
bf_M07  <-  as.matrix(b_M07_bac_family_df)

b_M08_bac_family_df  <- bacteria_family_df[,c(14)]
bf_M08  <-  as.matrix(b_M08_bac_family_df)

b_M09_bac_family_df  <- bacteria_family_df[,c(8)]
bf_M09  <-  as.matrix(b_M09_bac_family_df)

b_M10_bac_family_df  <- bacteria_family_df[,c(9)]
bf_M10  <-  as.matrix(b_M10_bac_family_df)



## ============================================================================================================ ##
## 'weights' can be negative
## So try to solve giving only NON-NEGATIVE weights via non-negative least-squares (NNLS)
## see: 'nnls'  from  https://cran.r-project.org/web/packages/nnls/nnls.pdf
## ============================================================================================================ ##


## ============================================================================================================ ##
## FIRST: Inspect the 'Residual Sum of Squares' [RSS] values
## ============================================================================================================ ##

soln_M01 <- nnls(bacteria_family_matrix_A,bf_M01)
soln_M02 <- nnls(bacteria_family_matrix_A,bf_M02)
soln_M03 <- nnls(bacteria_family_matrix_A,bf_M03)
soln_M04 <- nnls(bacteria_family_matrix_A,bf_M04)
soln_M05 <- nnls(bacteria_family_matrix_A,bf_M05)
soln_M06 <- nnls(bacteria_family_matrix_A,bf_M06)
soln_M07 <- nnls(bacteria_family_matrix_A,bf_M07)
soln_M08 <- nnls(bacteria_family_matrix_A,bf_M08)
soln_M09 <- nnls(bacteria_family_matrix_A,bf_M09)
soln_M10 <- nnls(bacteria_family_matrix_A,bf_M10)

solution <- cbind(soln_M01, soln_M02, soln_M03, soln_M04, soln_M05, soln_M06, soln_M07, soln_M08, soln_M09, soln_M10)

solution[2,]                ### Row 2 holds the values for the 'deviance', or 'Residual Sum of Squares' [RSS] values,
              ### that is the 'distance' of the solution vector from the projected vector, b



## ============================================================================================================ ##
## SECOND: Grab the solution 'weight' values, or 'X' from [A.X = b]
## pass matrix 'A' and each vector 'b' in turn to the NNLS function
## ============================================================================================================ ##

library("limSolve")
packageVersion("limSolve")

soln_M01 <- nnls(bacteria_family_matrix_A,bf_M01, verbose = TRUE)
soln_M02 <- nnls(bacteria_family_matrix_A,bf_M02, verbose = TRUE)
soln_M03 <- nnls(bacteria_family_matrix_A,bf_M03, verbose = TRUE)
soln_M04 <- nnls(bacteria_family_matrix_A,bf_M04, verbose = TRUE)
soln_M05 <- nnls(bacteria_family_matrix_A,bf_M05, verbose = TRUE)
soln_M06 <- nnls(bacteria_family_matrix_A,bf_M06, verbose = TRUE)
soln_M07 <- nnls(bacteria_family_matrix_A,bf_M07, verbose = TRUE)
soln_M08 <- nnls(bacteria_family_matrix_A,bf_M08, verbose = TRUE)
soln_M09 <- nnls(bacteria_family_matrix_A,bf_M09, verbose = TRUE)
soln_M10 <- nnls(bacteria_family_matrix_A,bf_M10, verbose = TRUE)

solution <- cbind(soln_M01$X, soln_M02$X, soln_M03$X, soln_M04$X, soln_M05$X, soln_M06$X, soln_M07$X, soln_M08$X, soln_M09$X, soln_M10$X)


## name columns and rows ##
```

```
#########################
dimnames(solution) = list( c("P1","P4","P5","P8","P9","P10","P12","P13","P15"),c("M1","M2","M3","M4","M5","M6","M7","M8","M9","M10") )

solution  ## view the solution 'weights' for each mixed sample

write.table(solution, sep="\t", "solutionWeights.txt")


## ================================================================================================== ##
## THIRD: plot out the 'weights'
## Plot the weight of contribution for each of the single 'seed' samples to a mixture
## ================================================================================================== ##


## view weights for each mixed sample as barcharts  ##
## munge solution vectors into one vector           ##
## the following was adapted from: http://www.r-bloggers.com/using-the-svd-to-find-the-needle-in-the-haystack/ ##


library(lattice)
b_clr <- c("steelblue", "darkred")
b1 <- barchart(as.table(solution[,1]),
            main="M_01",
            horizontal=FALSE, col=ifelse(solution[,1] > 0,
                                b_clr[1], b_clr[2]),
            ylab="Weight",
            scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
            ) ###key = key)
    print(b1, split=c(1,1,3,4), more=TRUE)    ### 'split' is used to lay out the barchart lattice
                                                ### e.g. split=c(1,1,3,2) means place plot b1 in col. 1, row 1

b2 <- barchart(as.table(solution[,2]),
             main="M_02",
             horizontal=FALSE, col=ifelse(solution[,2] > 0,
                                b_clr[1], b_clr[2]),
             ylab="Weight",
             scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
             ) ###key = key)
    print(b2, split=c(2,1,3,4), more=TRUE)              ### 'split' is used to lay out the barchart lattice
                                                        ### e.g. split=c(2,1,3,2) means place plot b2  in col. 2, row 1
                                                        ### where the layout has 3 columns and 2 rows        x O x
                                                        ###                                                x x x

b3 <- barchart(as.table(solution[,3]),
             main="M_03",
             horizontal=FALSE, col=ifelse(solution[,3] > 0,
                                b_clr[1], b_clr[2]),
             ylab="Weight",
             scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
             ) ###key = key)
    print(b3, split=c(3,1,3,4), more=TRUE)


b4 <- barchart(as.table(solution[,4]),
             main="M_04",
             horizontal=FALSE, col=ifelse(solution[,4] > 0,
                                b_clr[1], b_clr[2]),
             ylab="Weight",
             scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
             ) ###key = key)
    print(b4, split=c(1,2,3,4), more=TRUE)


b5 <- barchart(as.table(solution[,5]),
             main="M_05",
             horizontal=FALSE, col=ifelse(solution[,5] > 0,
                                b_clr[1], b_clr[2]),
             ylab="Weight",
             scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
             ) ###key = key)
    print(b5, split=c(2,2,3,4), more=TRUE)


b6 <- barchart(as.table(solution[,6]),
             main="M_06",
             horizontal=FALSE, col=ifelse(solution[,6] > 0,
                                b_clr[1], b_clr[2]),
             ylab="Weight",
             scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
             ) ###key = key)
    print(b6, split=c(3,2,3,4), more=TRUE)



b7 <- barchart(as.table(solution[,7]),
             main="M_07",
             horizontal=FALSE, col=ifelse(solution[,7] > 0,
                                b_clr[1], b_clr[2]),
             ylab="Weight",
             scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
             ) ###key = key)
    print(b7, split=c(1,3,3,4), more=TRUE)


b8 <- barchart(as.table(solution[,8]),
             main="M_08",
```

```
            horizontal=FALSE, col=ifelse(solution[,8] > 0,
                                         b_clr[1], b_clr[2]),
            ylab="Weight",
            scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
            ) ###key = key)
    print(b8, split=c(2,3,3,4), more=TRUE)


b9 <- barchart(as.table(solution[,9]),
            main="M_09",
            horizontal=FALSE, col=ifelse(solution[,9] > 0,
                                         b_clr[1], b_clr[2]),
            ylab="Weight",
            scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
            ) ###key = key)
    print(b9, split=c(3,3,3,4), more=TRUE)


b10 <- barchart(as.table(solution[,10]),
            main="M_10",
            horizontal=FALSE, col=ifelse(solution[,10] > 0,
                                         b_clr[1], b_clr[2]),
            ylab="Weight",
            scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
            ) ###key = key)
    print(b10, split=c(1,4,3,4))
```

Method S1: **Code for the Non-Negative Least Squares (NNLS) analysis related to STAR Methods.** This annotated code can be run using R in order to obtain the contribution of individual communities towards the community mix. The input files needed are the .biom file containing the compositions of communities analysed and their phylogeny, preferably prepared using Cumulative Sum Scaling normalisation (see STAR methods). It also requires a mapping file as described in http://qiime.org/documentation/file_formats.html. The version of the code is suited for analysis of our dataset but can be readily adapted for any dataset containing multiple communities amplicon data.

The script was run in RStudio v1.0.143 using R version 3.4.0 (2017-04-21) and the following R package versions: limSolve_1.5.5.2, nnls_1.4, biomformat_1.4.0, Phyloseq_1.20.0.